Machine Translation on a parallel Code-Switched Corpus

M.A. Menacer, D. Langlois, D. Jouvet, D. Fohr, O. Mella, and K. Smaïli

LORIA, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy, France

Abstract. Code-switching (CS) is the phenomenon that occurs when a speaker alternates between two or more languages within an utterance or discourse. In this work, we investigate the existence of code-switching in formal text, namely proceedings of multilingual institutions. Our study is carried out on the Arabic-English code-mixing in a parallel corpus extracted from official documents of United Nations. We build a parallel code-switched corpus with two reference translations one in pure Arabic and the other in pure English. We also carry out a human evaluation of this resource in the aim to use it to evaluate the translation of code-switched documents. To the best of our knowledge, this kind of corpora does not exist. The one we propose is unique. This paper examines several methods to translate code-switched corpus: conventional statistical machine translation, the end-to-end neural machine translation and multitask-learning.

Keywords: Code-switching · Machine translation · Statistical Machine Translation · Neural Machine Translation

1 Introduction

Code-switching (CS) can be defined as the use of more than one language by a speaker within an utterance. This phenomenon occurs generally in multilingual communities where speakers are known for their ability to code switch their languages during the communication.

CS is used both in informal (tweets and online content) and formal (proceedings of multilingual institution) texts. This makes the code-switched data a challenging issue for Natural Language Processing (NLP). Even if there were several linguistic studies on mixed languages [11,4], the computational processing of this kind of data remains relatively weak [2,10].

One of the challenges of the NLP concerning code-switched texts is the lack of data and tools. That is why, our first objective is to provide a code-switched dataset. Code-switched corpora are either generated in an artificial way [13,12] or collected from social media and/or online texts [1]. Once the data are collected, the processing of code-switching is carried out by adapting the existing models and tools or by proposing new ones.

Recent efforts related to CS are more about data collection and analysis [14,7]. Few works have explored downstream tasks as Language Modeling (LM) [8], Automatic Speech Recognition (ASR) [12] and Machine Translation (MT) [5].

In this article, we focus on the machine translation of Arabic-English code-switched documents. This is a challenging task for two reasons. In fact, the development of a MT system needs parallel corpora. This kind of data is usually available for pure languages, however, for mixed languages, there is no available resource. Thus, our first objective is to build a code-switched corpus with two reference translations, one for each language. This resource will mainly be used for tuning and testing the performance of the translation system. The second reason that makes difficult to translate code-switched documents is that models do not exist for this kind of data. Our approach in this work consists of using and adapting existing models in the aim to study the impact of the code-switched data on the machine translation.

2 How to build a code-switched parallel corpus?

Our main objective is to translate mixed Arabic-English documents to pure Arabic and/or to pure English forms. To reach this objective, we decided to investigate the United Nations (UN) documents, because they mix English with Arabic in the Arabic official documents.

The parallel Arabic-English corpus extracted from UN documents corresponds to the period between January 2000 and September 2009 (MultiUN [6]). Table 1 sets out some statistics about this corpus after tokenization, truecasing and cleaning the Arabic and English corpora.

Language	**#sentences**	**#words**	**#unique words**
Arabic	9.7M	232.7M	690k
English		275.3M	388k

Table 1. Statistics about the parallel corpus.

The particularity of this corpus is that the Arabic sentences can include English segments but also segments from French, Spanish or other languages. All these segments are written in Latin script.

All code-switched sentences are extracted from the entire corpus (i.e. sentences containing words in Arabic characters and words in Latin characters). This leads to a parallel code-switched corpus representing only 3% from the original corpus. Among these mixed sentences, we kept only those without URL links, email addresses and those that do not contain acronyms, since acronyms are generally not translated and are kept as they are.

Due to the parallel nature of the corpus, the English translation of each code-switched sentence can be easily recovered. For these remaining code-switched sentences, obviously Arabic is the dominant language. However, words in Latin script represent 12% of the total number of this corpus.

Code-switched sentences can be translated into pure English (for an English-speaking people) or pure Arabic (for an Arabic-speaking people). The English translation is already available but we do not have a pure translation into Arabic. To do so, we propose to produce it automatically. All the Arabic segments in a code-switched sentence are kept as they are and the English segments are translated into

Arabic by using Google translate API [1]. The addition of the Arabic segments and the translated English segments into Arabic constitutes a Pure Arabic Sentence, named *PAS*.

Table 2 describes some statistics about the resulting parallel code-switched corpus.

Language	**#sentences**	**#words**	**#unique words**
CS	37k	1.09M	99k
English ref		1.14M	64k
Arabic ref		1.06M	88k

Table 2. Statistics about the parallel code-switched corpus.

An example of a code-switched sentence and its reference translations (pure Arabic and pure English) is set forth in Table 3.

Code-switched sentence	*AmA bAl<nklyzyp fhy mrkbp mn Al>Hrf Al>wlY mn AlklmAt AltAlyp* boosting and inspiring dynamic youth achievement
Arabic reference translation	*AmA bAl<nklyzyp fhy mrkbp mn Al>Hrf Al>wlY mn AlklmAt AltAlyp* tEzyz w<lhAm Al<njAz AldynAmyky ll$bAb
English reference translation	It is also an acronym for Boosting and Inspiring Dynamic Youth Achievement

Table 3. Examples of a code-switched sentence with their Arabic/English translation reference (Arabic segments are written in buckwalter transliteration).

The Arabic reference translation is a mix of human produced data (original Arabic sentences) and Google machine translation produced data (the CS parts are translated back into Arabic); hence one needs to ensure the quality of the generated translation. That is why, we decided to evaluate manually the Arabic reference translation.

To investigate the effect of different sources of information on the evaluation procedure, we asked 6 people to evaluate 1200 sentences by using the five-point scale shown in Table 4 according to two evaluation scenarios:

Scenario 1: target only the user evaluate only *PAS* without any information about the source code-switched sentence.

Scenario 2: source+target participant has both *PAS* and the source code-switched sentence.

	Scenario 1	**Scenario 2**
1	incomprehensible	no relation between source and target
2	some segments are understandable	some segments are correctly translated
3	understandable but non-native Arabic	understandable translation but non-native Arabic
4	very understandable	understandable translation
5	excellent	good translation

Table 4. Evaluation scale used for each evaluation scenario.

The final score is calculated by computing the sample mean of all sentence judgments. The results are presented in Table 5.

Participants	p1	p2	p3	p4	p5	p6	**Mean**
Scenario 1	2.86	3.14	3.21	3.32	3.49	3.51	3.26
Scenario 2	4.09	4.10	4.53	4.04	3.53	4.00	4.05

Table 5. Average score of human evaluation of the Arabic reference translations.

[1] https://github.com/ssut/py-googletrans

Concerning the first scenario, all participants except *p1* consider globally that the produced *PASs* are understandable but suffer from a bad style in the form of the Arabic structure. For the second scenario, the participants consider that the produced *PASs* are understandable. This is probably due the influence of the users by the source sentence, which is was available in the second evaluation scenario.

For further information about how the participants have evaluated the translations, we decided to measure the inter rater reliability. To do so, participants should evaluate the same set of sentences, and thus another set of 100 sentences are selected randomly from the entire corpus and they are evaluated by each participant. The inter rater reliability is measured by using Fleiss' kappa coefficient [9]. The obtained value was 0.41 with a standard error of 0.02 which can be interpreted as there is a moderate agreement between participants. This is an expected result since the two categories *very understandable / excellent* in scenario 1 and *understandable translation / good translation* in scenario 2 are so close, hence for some sentences, this could easily be confused.

3 The impact of code-switched data on machine translation

In this section, we present several scenarios of machine translation. For each of them, we investigate both the conventional Statistical Machine Translation (SMT) and the Neural Machine Translation (NMT) approaches. The four scenarios are the following:

Baseline system With 1M of parallel sentences (pure Arabic and pure English), we trained machine translation systems and we tested them on a code-switched test corpus.

With output copy (WOC) system The baseline system is used except that the English segments in the code-switched test corpus are directly copied into the output.

Translation based on multilingual training corpus In this case, we trained a translation system with a code-switched corpus built automatically by replacing, in the Arabic corpus, some segments by their English translation. The translations are extracted from the phrase table of the baseline system.

Multitask learning Since the source corpus to translate is a mixture of Arabic and English segments, our idea is to train one model that performs translation in the two directions, from Arabic to English and vice versa.

Experiments were performed on 5k code-switched sentences extracted from our resource produced in Section 2. While the reference translation is available for both languages, the translation is just carried out from code-switched sentences to pure English. Table 6 sets forth the evaluation of translation in terms of BLEU metric and the Out-of-vocabulary rate (OOV).

Attempts	**SMT**	**NMT**	**OOV (%)**
Baseline	29.89	24.09	14.09
WOC	31.06	**33.11**	14.09
Multilingual	**32.06**	31.07	5.67
Multitask	/	28.36	14.09

Table 6. The evaluation of the machine translation systems

In the case where the test corpus is translated into pure English by using the baseline system, all segments with Latin script are considered as out-of-vocabulary; this explains the OOV rate of 14.09%. Besides, the NMT approach performs poorly on code-switched data compared to the SMT approach. This is due to the OOV rate. In fact, NMT performs on constraint vocabulary, all words that do not exist in the vocabulary are replaced by a special token. Even by substituting unknown words with source words that have the highest attention weight, the translation is not improved. One explanation for this would be that the attention model does not functionally play the role of a word alignment between the source and the target sentences. It provides only a soft-alignment to help the decoder to decide parts of the source sentence to pay attention to [3].

For this reason, preventing the translation of foreign segments, and copying them directly into the output (WOC in Table 6) improves the translation comparing with the baseline system and specially in the NMT approach.

Training the translation systems on an artificial code-switched corpus (multilingual in Table 6) improves the translation compared with the baseline system where we used a monolingual parallel corpus. This approach outperforms also the second system (WOC), where segments in foreign language are not translated but copied directly into the output. These improvements are justified by the decrease of the OOV rate (5.67% against 14.09%) and by the fact that the training corpus is approaching the test corpus. The same observation holds true for the NMT approach, except that training the neural network on an artificial multilingual corpus does not outperform the second system.

Ultimately, training one model that performs the translation in the two directions (multitask learning) yields a gain of 4% BLEU points over the baseline system where a single task is learned.

4 Conclusion

The study carried out in this work focused on the impact of code-switching on the translation system. We firstly investigated whether the code-switching occurs in formal text. We accomplished this through the Arabic-English parallel corpus extracted from the official documents of United Nations. From this corpus, we built and evaluated a parallel code-switched resource, which is available for free access[2]. It also provides a valuable resource for studying multilingual practices in other works.

Several training/translation strategies were investigated for both the SMT and the NMT approaches. Results showed that the conventional SMT reaches best performance if a multilingual model is trained on mixed languages. Besides, we found that avoiding translation of segments in foreign language is the best strategy for the end-to-end model. We also found that training the neural network on two translation tasks instead of one improved the translation of code-switched sentences.

Furthermore, in our work, the whole code-switched sentence is translated directly with the different systems; it would be interesting to identify implicitly the language

[2] `https://smart.loria.fr/Fichiers/MTCS.rar`

of the foreign segments and carry out the translation with an appropriate translation system as it was carried out in [13] for the recognition of mixed speech.

References

1. Abidi, K., Menacer, M.A., Smaïli, K.: Calyou: A comparable spoken algerian corpus harvested from youtube. In: 18th Annual Conference of the International Communication Association (Interspeech) (2017)
2. Abidi, K., Smaïli, K.: An empirical study of the Algerian dialect of Social network. In: ICNLSSP 2017 - International Conference on Natural Language, Signal and Speech Processing. Casablanca, Morocco (Dec 2017), `https://hal.inria.fr/hal-01659997`
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
4. Bullock, B.E., Hinrichs, L., Toribio, A.J.: World englishes, code-switching, and convergence. Oxford Handbook of World Englishes (2014)
5. Carpuat, M.: Mixed language and code-switching in the canadian hansard. In: Proceedings of the first workshop on computational approaches to code switching. pp. 107–115 (2014)
6. Eisele, A., Chen, Y.: MultiUN: A multilingual corpus from united nation documents. In: Tapias, D., Rosner, M., Piperidis, S., Odjik, J., Mariani, J., Maegaard, B., Choukri, K., Chair), N.C.C. (eds.) Proceedings of the Seventh conference on International Language Resources and Evaluation. pp. 2868–2872. European Language Resources Association (ELRA) (5 2010)
7. Gambäck, B., Das, A.: Comparing the level of code-switching in corpora. In: Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (may 2016)
8. Garg, S., Parekh, T., Jyothi, P.: Dual language models for code mixed speech recognition. CoRR **abs/1711.01048** (2017), `http://arxiv.org/abs/1711.01048`
9. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977)
10. Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., Solorio, T.: Overview for the second shared task on language identification in code-switched data. In: Proceedings of the Second Workshop on Computational Approaches to Code Switching. pp. 40–49 (2016)
11. Poplack, S.: Sometimes ill start a sentence in spanish y termino en espanol: toward a typology of code-switching1. Linguistics **18**(7-8), 581–618 (1980)
12. Toshniwal, S., Sainath, T.N., Weiss, R.J., Li, B., Moreno, P., Weinstein, E., Rao, K.: Multilingual speech recognition with a single end-to-end model. arXiv preprint arXiv:1711.01694 (2017)
13. Watanabe, S., Hori, T., Hershey, J.: Language independent end-to-end architecture for joint language identification and speech recognition pp. 265–271 (12 2017)
14. Yoder, M., Rijhwani, S., Rosé, C., Levin, L.: Code-switching as a social act: The case of arabic wikipedia talk pages. In: Proceedings of the Second Workshop on NLP and Computational Social Science. pp. 73–82 (2017)